Signals Intelligence - Processing - Analysis - Classification

Ulla Uebler
Business Development
MEDAV GmbH
Uttenreuth, Germany

Ulla.Uebler@medav.de

Hans-Joachim Kolb
General Management
MEDAV GmbH
Uttenreuth, Germany

Hans-Joachim.Kolb@medav.de

ABSTRACT

In the world of SIGINT / COMINT more data are incoming from a large variety of sources – for example mobile and satellite communication. Automatic systems become necessary to process the amounts of data.

In this paper we focus two main aspects: (1) the treatment and processing of data coming from different sources in the same standard manner, i.e. all inputs formats will be treated in the same way, leading to a standardized data model; and (2) introduce automatic processing wherever possible – applying either manual or automatic algorithms where they are best suited.

The way of applying these two procedures is firstly realised in our SIPAC system to be described later on in more detail. We have developed this system in years of applied research; the technical algorithms for automatic processing are mainly developed at MEDAV – although the system is flexible enough to integrate specialized algorithms from other parties.

It is a key item of this paper to show the effort spent while researching the optimal performance of such a system still leaving the flexibility to humans to modify the parameters according to a specific task.

Keywords: Intelligence, COMINT, SIGINT, OSINT information fusion, speech classification, text analysis, image analysis, modular system architecture.

1.0 INTRODUCTION

In this paper we describe our particular approach to the processing of data coming in from different sources, different places and different types.

Main steps are:

- Gather data at the source,
- Add available meta-information, e.g. the name and type of source,
- Throw away data already classified as unimportant,
- Store data according to our data model in the system database, together with the meta-data,
- Depending on the current task/mission: start automatic processing using meta-information to enrich the data automatically,
- At any time an operator may manually do his analysis,
- Human experts are involved to control the process of automatic processing, e.g. for quality management, retraining and other aspects.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
OCT 2009	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Signals Intelligence - Processing - Analysis - Classification	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Business Development MEDAV GmbH Uttenreuth, Germany**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADB381582. RTO-MP-IST-087 Information Management - Exploitation (Gestion et exploitation des informations). Proceedings of RTO Information Systems Technology Panel (IST) Symposium held in Stockholm, Sweden on 19-20 October 2009., The original document contains color images.

14. ABSTRACT
In the world of SIGINT / COMINT more data are incoming from a large variety of sources for example mobile and satellite communication. Automatic systems become necessary to process the amounts of data. In this paper we focus two main aspects: (1) the treatment and processing of data coming from different sources in the same standard manner, i.e. all inputs formats will be treated in the same way, leading to a standardized data model; and (2) introduce automatic processing wherever possible applying either manual or automatic algorithms where they are best suited. The way of applying these two procedures is firstly realised in our SIPAC system to be described later on in more detail. We have developed this system in years of applied research; the technical algorithms for automatic processing are mainly developed at MEDAV although the system is flexible enough to integrate specialized algorithms from other parties. It is a key item of this paper to show the effort spent while researching the optimal performance of such a system still leaving the flexibility to humans to modify the parameters according to a specific task.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**SAR**	**8**	

A lot of effort has been accomplished by our experts in order to obtain the required information the quickest possible, with the least errors in the interpretation, thus obtaining an optimized information flow.

Since we want to analyze and evaluate signals form different sources and types in the same manner, we first need a data model that is well suiting to the required needs of the analyzing operators. We describe the data model we have developed for our system SIPAC in the next section.

In section 3.0 we describe the information flow in our system from the source of the signal via (automatic) processing, towards analysis and classification.

Section 4.0 describes a choice of available algorithms – these algorithms are usually provided by MEDAV, due to the flexible system architecture they may be provided as well as from the customer or from third parties.

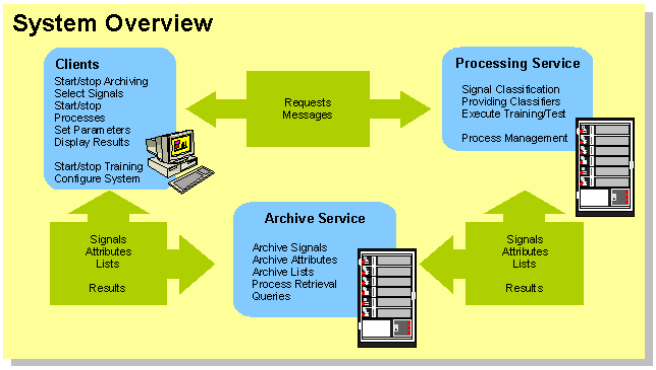

Figure 1: System Overview

The criterion for a good performing system is not only to have good working parts, but above all, to have a fruitful and efficient interaction between the parts.

2.0 DATA MODEL

The data model used in this scenario is one of the top three decisions to make the system work optimally. Only when having an optimal data model, it becomes possible to

- Adequately store the retrieved information
- Find the stored information in the huge data base in a reasonable amount of time.

This process of developing the optimal data model has been experienced at MEDAV during the last 20 years, and is right now at a very versatile state.

Current systems use different data models for each type of information. For example, for text messages, a variety of information becomes available by automatic processing (e.g. language identification, key word extraction) together with the results of manual analysis often added as a comment or report linked in some way to the original signal. A different set of parameters is used for wideband signals, e.g. direction or location of an emitter, technical parameters like the coding of signals, the frequency range of signals in the wideband signal etc.

With our system SIPAC we have elaborated a modelling of attributes that describe information and analysis results in a comparable manner in a way that analysis results of different sources can be compared. The attributes representing the information with respect to a given document in a large

structure, containing more than 100 different types of attributes; of course together with an elaborated search mechanism that may search in all or specified types of attributes.

The basic idea is that we treat any signal delivered to the system as a file. An additional meta-file is attached to each file containing attributes and additional information. During the process of analysis, obtained information is added to the meta-file.

Information can be entered either via

- Manual analysis by specialists, using their background knowledge in either filling out a standard form or in a so-called free text field.
- Automatic analysis using the algorithms integrated in the SIPAC system, e.g. for text classification, that fills in the respective fields in the meta-file standard form.

Basic form entries to be filled are

- Date/time stamp of incoming signal
- Type of signal (text, speech, wideband signal)
- Size of file
- Name or identification of source.
- Free text field to be filled by operators.

Other fields can be more specific depending on the type of the signal. For text files, this could be

- Language
- Topic
- Summary
- Translated text
- Keywords
- …and a lot more

Using this type of data model, we are able to store the information in a structured way. The operator is now able to perform efficient search only in the meta-data files with a fractional amount of storage size and thus of search time.

3.0 INFORMATION FLOW

The architecture of the information flow is another item of the top three to provide a well-functioning system. A very special point in the flow of processing the incoming information towards a condensed information structure is the combination especially of automatic and manual analysis, which optimally works hand in hand for the best system performance.

MEDAV has developed the information flow in a series of research projects, and is able to maintain the information flow on a widespread area of applications.

A typical way of analysing mass data from the source to the final report and/or eventually warning message is often performed in three steps, described in the following. Our system IFS-8000 treats the

complete chain including data fusion, automatic evaluation, manual analysis processed e.g. by jobs from a supervisor to the operators until finally a report is generated for the complete mission

3.1 Pre-processing

Gathered data are pre-processed directly of the sensor. Some signals are thrown away because there is evidence that they are not important for the respective organisation. Example: in the field of radio monitoring, certain emitters may not be in the range of interest, since they are located in a region of out of interest.

During pre-processing, some basic attributes are added to the meta-information file. All remaining files are passed to the SIPAC system for automatic classification.

3.2 Automatic classification

The SIPAC system has a set of classification algorithms defined by the supervisor and/or specialists of an organisation. The classifiers themselves are described in the following section.

All data coming from different sources arrive at the SIPAC system and are automatically processed depending on their type and pre-defined signal flows.

These signal flows are configured by system experts and can be modified with changing needs. An example simple signal flow:

FOR TEXT FILES	
ALG: Language identification	
-> ENGLISH	-> FRENCH
ALG: English keyword finder	*ALG: French keyword finder*
	ALG: Keyword translation to English

Depending on the interest of the operators and the included algorithms, different classifiers can be used in any sequence.

The information gathered in this process, is stored automatically in the meta-information file and can be used in the following manual analysis.

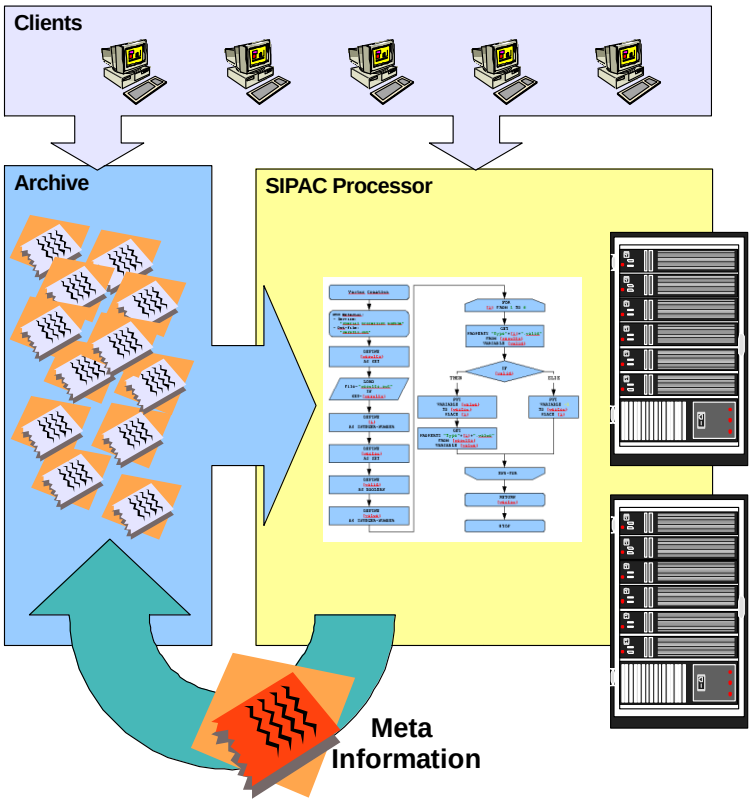

Figure 2: Adding Meta Information

Figure 2 describes the way of processing. Starting on the left side, we find all data in the data base archive. They can be processed using a defined processing chain, adding the meta-information to the data base archive again. The operators can modify and interact both with the data base and the SIPAC processing unit to perform the necessary tasks.

3.3 Manual processing

In the processing chain, the third step is the manual analysis by human operators and experts. The operators perform their analysis in different ways depending on the current job, e.g.

- Statistical analysis in the meta-information file.
- Search specific entities (names, persons, places in the meta-information and in the data themselves.
- Write reports which may be stored again as meta-information or as new documents (eventually again entering in the processing chain e.g. with using analysis).

3.4 Operator roles in the system

Human experts (for different things, we will see later) have a very important role in the functioning of a system. The role of SIPAC is to provide a platform to the experts where they obtain support in fulfilling their tasks, quickly finding the required information, related documents etc.

Example: Language identification from audio signals.

In a certain mission, a set of languages seems important beforehand. These languages will – with a certain percentage of failure – be recognized by the system automatically. Other languages, rated as not typical to the mission, will not be used in the automatic classification process.

An analysis expert will notice that some languages have a higher misclassification. There he can notify other experts on that:

- The translation expert for the rare and in this case misclassified language – for a fast translation of the special event
- The linguist expert in order to have him check on the parameters of the language identification algorithm – maybe a retraining of the parameters may become necessary.

As this example shows, it is very important, that the experts are capable to quickly pass information to a completely different type of experts, leading to a fast improvement of the system in case of changing parameters. MEDAV provides in these notices all necessary information for the other experts to have them easily find the necessary information (e.g. the speech file to be translated by humans).

Another important point consists of the implementation of the "Intelligence Cycle" as described by MEDAV: the result of a mission is composed of a report and a possible reaction. If, caused by the report, a new mission is started, the information from that mission may be used directly, e.g. trained algorithms, basic knowledge.

MEDAV investigated if the training and optimizing of algorithms should be a fixture to the system. After evaluation, we must strictly say "yes", since this is the most direct method to obtain and retain the best possible performance. What looks at first sight as a withdraw, is the need of algorithm experts somewhere inside the system. But according to our data model the expert can be geographically be at a different place and may even be in charge of algorithms in a set of missions – depending on the system allowances given by the system administrator.

4.0 CLASSIFICATION ALGORITHMS

The deep understanding of involved algorithms is the third of the top three features of a optimally working system. MEDAV has developed in many years algorithms for speech, text, and image classification, as well as for signal processing in general (also inside the radio monitoring domain).

This experience allows a reliable judgement of the manner and place where these algorithms should be applied. MEDAV has algorithms for almost all classifications described below, but the system architecture is kept open in order to allow the customer to use his own or other algorithms, also from other domains, e.g. cryptology, traffic analysis or geographic information systems – the data model is open to store and retrieve also this special information.

When we think of information processing as a three-step process, automatic classification algorithms can be used in the first and second step.

Automatic classification algorithms serve for a fast processing of many data. These automatic algorithms have a typical error rate, depending on the type of algorithm and on the difficulty of the task. As an example, speech detection is a simpler algorithm than language identification and thus yields typically a lower error rate than language identification.

Some of the algorithms have to be trained before using them. During training, a set of parameters is estimated using labelled signals of the scenario, i.e. using the known signals as typical examples and thus setting the processor accordingly.

A typical first step in automatic processing is the detection of the signal type, i.e. to determine if the incoming signal is wideband, narrowband or of other types. After this classification, the algorithms that suit to the type of signal will be used. Typical classification algorithms are listed in the following.

4.1 Algorithms Speech

A choice of algorithms is available. Some of the algorithms have quite good recognition accuracy, others depend strongly on the channel conditions and may thus not provide good accuracy for disturbed signals.

- Speech detection: determines speech parts in a signal compared to silence, noise – algorithm trainable or using thresholds.
- Language identification: determines the spoken language in a speech signal. Trainable algorithm – the languages to be recognized must be trained with suitable data.
- Speaker identification: determines a certain speaker in a speech signal. Algorithms do a time/file-wise classification or based on speech segments. Trainable algorithm – the speakers to be recognized must be trained before.
- Topic spotting: determines the topic/theme in a speech signal, e.g. politics, sports. Trainable algorithm – the topics to be recognized must be trained.
- Word spotting: detects words out of a pre-defined word list. Trainable algorithm – the sounds of the respective language must be trained, the words to be determined must be known to the system with respect to their sounds.
- Transcription of speech signals: transcribes the utterance in a speech signal word by word. Trainable algorithm – the sounds of the language must be trained, the words to be recognized must be represented by the sounds of the language.

4.2 Algorithms Images

Also here, a choice of algorithms is available, depending on the tasks to be performed.

- OCR: determine the text parts in an image – language dependent approach, quality depends on the language.
- Steganography: detects hidden text in images – trainable algorithm
- Detection of objects and persons in an image. Trainable algorithm – the object or person to be detected must be known to the system before.

4.3 Algorithms Text

Text is somehow the final form of signal and information processing and analysis. Speech and image information is stored as text. Reports are written as text. Therefore a set of algorithms is available to automatically process text.

- Language identification: determines the language a text is written in. Trainable algorithm – the languages to be recognized must be known to the system.
- Key word detection: detection of either pre-defined word list or special word types like persons and places – trainable algorithm.

- Text summary: selection of the most important parts of a text – trainable algorithm. The automatic generation of a text summary is still a difficult task.
- Text translation: translation of foreign language texts to the main language. Full and correct translation is subject to research, translation of text phrases is already feasible and helpful.

4.4 Special Algorithms

Any special algorithms – user defined, customer-owned can be integrated into the SIPAC system. These may be special decryption algorithms or any software of the users.

5.0 SUMMARY

We have presented a system for flexible and modular processing of any types of signals. The advantage of such an open system lies in the flexible structure of the database using meta-information to describe the information gathered in the processing steps – supported by automatic algorithms. All the information is gathered in the database in the same storage philosophy which makes it easy to retrieve information originating from different source types.

A large variety of processing tools can be embedded in this system using its modular structure.

According to our long experience in this field we know of the top three factors for a successful processing system:

- A data model, that is open for the storage of specific and detailed information. A crucial point of the data model is to quickly find the required information – a good organisation of the attributes is necessary – and provided by MEDAV.

- A deep understanding of the information flow and the underlying processes of analysis – and the strong interaction between automatic and manual analysis

- A set of qualified algorithms as well as the knowledge to judge on the quality and meaning of automatic algorithms.

Last argument to point out is that the system presented here, relies on many years of experience in these fields – the total system is a lot more than the simple sum of its parts – these parts must rely on each other and interact in a meaningful and efficient manner.

REFERENCES

[1] Kolb, Hans-Joachim and Towsey, Michael W. and Maetschke, Stefan and Uebler, Ulla (2003) The Visualisation of Diverse Intelligence. In Proceedings NATO (Research and Technology Agency) conference on "Military Data and Information Fusion. Paper 23, Prague, Czech.

[2] Thurmair, Gregor (2005): Hybrid Architectures for Machine Translation Systems. in: Language Resources and Evaluation 39, 1, 91-108.

[3] Uebler, Ulla (2006) "A Speech Classification System". In Information Fusion for Command Support (pp. 12-1 – 12-10). Proceedings RTO-MP-IST-055, Paper 12. Neuilly-sur-Seine, France: RTO.

[4] Uebler, Ulla (2001), "Multilingual speech recognition in seven languages," in Speech Communication, Aug. 2001, pp. 53–69.